

TP 2: Análisis de Redes Sociales

Octubre 27, 2022

Métodos Numéricos

Grupo 18

Integrante	LU	Correo electrónico
Vekselman, Natán	338/21	natanvek11@gmail.com
Arienti, Federico	316/21	fa.arianti@gmail.com
Manuel Lakowsky		
Brian Kovo		



Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
<http://www.exactas.uba.ar>

RESUMEN

La descomposición de matrices en autovectores y autovalores aparece en una variedad de aplicaciones donde importa caracterizar el comportamiento de un sistema: en el reconocimiento de imágenes, en el análisis de estabilidad de cuerpos rotantes, en el análisis de riesgo de mercado y en el análisis de redes —por nombrar algunas—. Desde un punto de vista geométrico, se puede considerar a los autovectores como los ‘ejes’ de una transformación lineal, en tanto representan una dirección invariante a la transformación, y a los autovalores como los factores por los que esas direcciones se comprimen, estiran o invierten.

En este trabajo propondremos una implementación en C++ de un método para el cálculo de autovalores y autovectores en matrices cuadradas. Para algunas matrices particulares, como pueden ser las matrices simétricas definidas positivas, este método nos permitirá obtener todos sus autovalores y autovectores asociados. El mismo se conoce como *el método de la potencia con deflación*.

A su vez, presentaremos dos aplicaciones concretas de los autovalores y autovectores en el análisis de redes: la medición de centralidad de autovector y corte mínimo en la red del ‘Club de Karate’ [6], y la estimación de una *ego-network* [4] de Facebook, por medio de la construcción de una matriz de similaridad.

Palabras clave: *método de la potencia, deflación de Hotelling, centralidad de autovector, conectividad algebráica, análisis de componentes principales.*

CONTENIDOS

1. Método de la potencia con deflación	2
1.1. Introducción teórica	2
1.2. Implementación	3
1.3. Evaluación cuantitativa	5
2. Análisis: Club de Karate	8
2.1. Contexto	8
2.2. Centralidad de Autovector	9
2.3. Autovectores de la matriz laplaciana	10
3. Análisis: Red ‘Ego’	14
3.1. Contexto	14
3.2. Matriz de similaridad	14
3.3. Comparación con la red original	16
3.4. Optimización	17
3.5. PCA	17
4. Conclusiones	18
5. Apéndice	19
Referencias	20

1. MÉTODO DE LA POTENCIA CON DEFLACIÓN

1.1. Introducción teórica. El método de la potencia con deflación permite aproximar un subconjunto de los autovalores y autovectores asociados a una matriz. Si la misma satisface que todos sus autovalores son no nulos y diferentes en módulo, entonces permite aproximar el conjunto entero.

MÉTODO DE LA POTENCIA. El método de la potencia, *Power method* o *Power iteration*, es una técnica iterativa para aproximar el autovector asociado al autovalor en módulo máximo de una matriz cuadrada que satisfaga esta característica —es decir, tenga un autovalor dominante no nulo—, a partir de la aplicación de sucesivos productos matriciales, descriptos por la siguiente relación de recurrencia:

$$b_0 \text{ es un vector aleatorio : } \|b_0\| = 1$$

$$(1) \quad b_{k+1} = \frac{\mathbf{A}b_k}{\|\mathbf{A}b_k\|}$$

donde $\|\cdot\|$ es una norma vectorial.

Se puede demostrar [2] que, bajo las condiciones descriptas, si b_0 no es ortogonal al autovector asociado al autovalor dominante en módulo de \mathbf{A} , b_k convergerá a éste. Lo que es más, se podrá aproximar el autovalor dominante por medio del coeficiente de Rayleigh:

$$(2) \quad \lambda_{max} = \frac{b_k^t \mathbf{A} b_k}{b_k^t b_k}$$

MÉTODO DE LA DEFLACIÓN. El método de la deflación, por su parte, corresponde a la transformación de la matriz inicial \mathbf{A} por una matriz \mathbf{B} con autovalores equivalentes, salvo por el autovalor dominante que será anulado. Existen distintos métodos de deflación, entre ellos la deflación de Hotelling y la deflación de Wielandt [2].

En este trabajo utilizaremos la deflación de Hotelling por su sencillez, a cuestas de un mayor error numérico [2]. El mismo consiste en aplicar el método de la potencia para sucesivas matrices que satisfagan la siguiente relación:

$$\mathbf{B}_0 = \mathbf{A}$$

$$(3) \quad \mathbf{B}_{k+1} = \mathbf{B}_k - \lambda_k v_k v_k^t$$

donde λ_k corresponde al k -ésimo autovalor en módulo máximo de \mathbf{A} , estimado por el método de la potencia, y v_k es su autovector asociado.

1.2. Implementación. Procedemos a detallar una posible implementación¹ para ambos métodos, restringiéndonos al caso de autovalores reales. Definimos:

deflacion : $\text{matriz}_{n \times n} \mathbf{A} \times \text{nat } k \times \text{nat } q \times \text{real } t \longrightarrow \text{vector}_k \times \text{matriz}_{n \times k}$

potencia : $\text{matriz}_{n \times n} \mathbf{A} \times \text{nat } q \times \text{real } t \longrightarrow \text{real} \times \text{vector}_n$

donde n es un natural, \mathbf{A} tiene al menos k autovalores reales dominantes en módulo, $0 < k \leq n$, q es un número par² que representa el máximo de iteraciones a realizar y $0 \leq t$ representa la tolerancia mínima a partir de la que se considera la convergencia de una solución.

```

1 proc deflacion(in A: matriz<n, n>, in k: Nat, in q: Nat, in t: Real) {
2
3     eigvals := vector<q>
4     eigvecs := matriz<n, k>
5
6     i := 0
7     while i < k {
8
9         l, V := potencia(A, q, t)
10        eigvals[i] := l
11        eigvecs.columna[i] := V
12
13        A := A - l * (V * V.t)
14        i := i + 1
15    }
16
17    return eigvals, eigvecs
18 }
```

ALGORITMO 1. Pseudocódigo para el método de la deflación.

El algoritmo (1.) retornará un vector con los primeros k autovalores en módulo máximos de \mathbf{A} , ordenados descendientemente, y una matriz cuyas columnas corresponden, respectivamente, a los autovectores normalizados asociados a estos autovalores.

Es interesante notar que la k -ésima matriz sobre la que se aplicará el método de la potencia — \mathbf{B}_k — tendrá, por definición, un autovalor cero con multiplicidad algebraica mayor o igual a k . En el caso en que la matriz inicial sea simétrica, \mathbf{B}_k será simétrica³ y, en consecuencia, diagonalizable. Se puede demostrar que la única matriz diagonalizable con multiplicidad algebraica $\mu_a(0) = n$ es la matriz nula, por lo que el método de la deflación de Hotelling

¹El código, propiamente, se encuentra en [./implementacion/src/](#).

²Esta restricción no es necesaria, pero permite realizar una optimización que se explicará en la próxima página.

³Por suma de simétricas. Notar que vv^t siempre resulta en una matriz de este tipo.

tenderá hacia ella. Esto nos permite inferir que el error numérico será proporcional a k^4 . Es decir, los autovalores más chicos de la matriz serán más susceptibles a errores.

Por su parte, el algoritmo (2.) retornará el autovalor de \mathbf{A} máximo en módulo y su autovector asociado:

```

1 proc potencia(in A: matriz<n, n>, in q: Nat, in t: Real) {
2
3     v := aleatorio(n)      // un vector aleatorio no nulo
4     v := v / norma(v, 2)   // ||v||_2 = 1
5
6     i := 0
7     while i < q {
8
9         y := A * v
10        y := y / norma(y, 2)
11        if (norma(v - y, 2) < t) { // tolerancia para la convergencia
12            break
13        }
14        v := y
15        i := i + 1
16    }
17
18    l := (v.t * A * v) / (v.t * v) // coeficiente de rayleigh
19
20    return l, v
21 }
```

ALGORITMO 2. Pseudocódigo para el método de la potencia *monte carlo*.

Observamos que el algoritmo (2.) no es capaz de distinguir si la selección inicial del vector v es ortogonal al autovector asociado al autovalor en módulo dominante, por lo que una mala selección puede resultar en una respuesta incorrecta. Este tipo de algoritmo —que depende de un proceso aleatorio— se denomina *monte carlo* [1].

Una variante posible, de tipo *las vegas*, se presenta en el algoritmo (3.). Hasta llegar a un resultado aceptable, determinado por ϵ , o superar la cantidad de iteraciones permitida, definida por α , el algoritmo volverá a intentar resolver el problema. De no alcanzar una respuesta aceptable retornará una señal de error.

De este modo, la probabilidad de fallar a causa de una selección inicial ortogonal al autovector al que se espera converger será inversamente proporcional a α^5 .

⁴En tanto existirá una correlación. Sin embargo, es esperable que otros factores entren en juego: la varianza de los autovalores, el número de condición de la matriz, la selección del vector aleatorio inicial, o la cantidad de iteraciones q a realizar, por ejemplo.

⁵Esto es, considerando que cada selección inicial es independiente.

Para reducir esta probabilidad desde el comienzo, se propone que se elija cada coordenada del vector de manera pseudo-aleatoria sobre un rango amplio⁶. Por ejemplo, la máxima representación de enteros con signo en 32 bits.

```

1 proc potencia'(in A: matriz<n, n>, in q: Nat, in t: Real,
2                 in alpha: Nat, in epsilon: Real) {
3
4     res := false
5
6     i := 0
7     do {
8
9         l, v := potencia(A, q, t)
10        res := norma(A * v - v * l, 2) < epsilon
11        i := i + 1
12
13    } while not res and i < alpha
14
15    if !res {
16        return -1
17    }
18    return l, v
19 }
```

ALGORITMO 3. Pseudocódigo para el método de la potencia *las vegas*.

Es importante mencionar que una mala selección de ϵ —en función de q y t — resultará en un algoritmo que siempre falla. Esto dependerá de la velocidad de convergencia del método para la matriz particular sobre la que se lo aplica.

1.3. Evaluación cuantitativa. Procedemos a realizar una evaluación de nuestra implementación en C++ acorde a los algoritmos propuestos⁷. Se optó por utilizar el método de la potencia *monte carlo*.

Medimos el error relativo $\|\mathbf{A}\bar{\mathbf{V}} - \bar{\mathbf{V}}\bar{\Lambda}\|$ y absoluto $\|\Lambda - \bar{\Lambda}\|$ para 300 instancias de matrices $\mathbf{A} \in \mathbb{R}^{20 \times 20}$ generadas aleatoriamente, donde $\bar{\mathbf{V}}$ y $\bar{\Lambda}$ representan, respectivamente, las matrices aproximadas de autovectores y autovalores de \mathbf{A} .

METODOLOGÍA. Se calculó $\bar{\Lambda}, \bar{\mathbf{V}} = \text{deflacion}(\mathbf{A}, 20, 2e^4, 1e^{-20})$ y se midió el error relativo y absoluto del resultado.

⁶Desde un punto de vista geométrico, dos vectores son ortogonales sólo si son perpendiculares. De manera intuitiva, podemos ver que a medida que la dirección inicial posible de un vector tiende al infinito, la probabilidad que forme un ángulo de 90 grados con otro tiende a cero.

⁷El script asociado se puede encontrar en `./experimentos/error-potencia.py`, las tablas resultantes en `./experimentos/resultados/error-potencia`.

Cada caso se generó a través de uno de los siguientes tres procedimientos⁸:

- 1) *Matrices Diagonales*: Se generaron cien matrices diagonales \mathbf{D} con cien autovalores en el rango $[-1e^3, 0] \cup (0, 1e^3]$, con paridad diferente acorde al signo⁹. Los mismos se generaron con el rng *PCG64* de numpy.
- 2) *Matrices Diagonalizables*: Se generaron cien matrices diagonalizables $\mathbf{A} := \mathbf{Q}\mathbf{D}\mathbf{Q}^t$ donde cada matriz \mathbf{D} se generó a partir de la metodología (1.) y $\mathbf{Q} := \mathbf{I} - 2uu^t$ se generó a partir de un vector aleatorio u —con el algoritmo *random.rand()* de numpy— tal que $\|u\|_2 = 1$.
- 3) *Matrices Simétricas Definidas Positivas*: Se generaron cien matrices simétricas definidas positivas de enteros $\mathbf{A} := \mathbf{B}\mathbf{B}^t$ donde \mathbf{B} es una matriz aleatoria generada con el algoritmo *random.randint()* de numpy para el rango $[-1e^3, 0] \cup (0, 1e^3]$.

OBSERVACIONES. Consideramos que la varianza de los autovalores, el número de condición de las matrices, y su tamaño, son variables que afectan de manera significativa en el error del procedimiento.

El proceso mencionado para la generación de matrices aleatorias fue pensado sobre generadores de números aleatorios para tratar de minimizar cualquier tendencia que pueda provenir de la utilización de distribuciones particulares. De esta manera se espera que la varianza de los autovalores y el número de condición de las matrices también sean aleatorios, tal que los resultados brinden un panorama amplio del dominio de aplicación del método propuesto.

Además, se controló el tamaño de n , la cantidad de iteraciones q y la tolerancia t . Un estudio más exhaustivo debería también evaluar el comportamiento del algoritmo en función de estos parámetros.

RESULTADOS. La Figura (1.) resume los resultados para el error relativo.

test	L_1	L_∞
mediciones	300	300
error relativo promedio	$2.488868e^{-3}$	$3.973441e^{-4}$
desviación estándar	$3.403934e^{-2}$	$5.849651e^{-3}$
mínimo	0.0	0.0
25%	0.0	0.0
50%	$1.741562e^{-11}$	$1.043354e^{-11}$
75%	$4.240635e^{-6}$	$7.861135e^{-7}$
máximo	$5.812401e^{-1}$	$1.008212e^{-1}$

FIGURA 1. Datos de resumen para el error relativo $\|\mathbf{A}\bar{\mathbf{V}} - \bar{\mathbf{V}}\bar{\Lambda}\|$.

⁸Se utilizó un valor semilla para facilitar la reproductibilidad.

⁹De esta manera se garantiza la dominancia estricta en módulo de los autovalores asociados a la matriz.

La Figura (2.), por su parte, resume los resultados para el error absoluto.

test	L_1	L_∞
mediciones	300	300
error relativo promedio	$3.490778e^{-7}$	$2.069772e^{-7}$
desviación estándar	$5.051277e^{-6}$	$3.309506e^{-6}$
mínimo	0.0	0.0
25%	0.0	0.0
50%	$2.130074e^{-12}$	$3.410605e^{-13}$
75%	$1.393337e^{-7}$	$3.352761e^{-8}$
máximo	$8.753557e^{-5}$	$5.733664e^{-5}$

FIGURA 2. Datos de resumen para el error absoluto $\|\Lambda - \bar{\Lambda}\|$.

Vemos que el metodo funcionó, para el 75% de las matrices evaluadas, con un error relativo y absoluto menor a $1e^{-4}$. Sin embargo, notamos que existen casos anomalos para los que el error puede ser mayor.

Por la disparidad entre el error relativo y absoluto, podemos inferir —a modo de futura hipótesis a investigar— que el método produce más error en los autovectores resultantes que en los autovalores.

2. ANÁLISIS: CLUB DE KARATE

2.1. Contexto. La red del *Club de Karate* fue parte de una investigación antropológica [6], realizada por Wayne W. Zachary, que estudió las relaciones ‘políticas’ entre los miembros de un club universitario. La misma se realizó durante el desarrollo de un conflicto que terminó por dividir al grupo.

La red buscó modelar el flujo de información entre sus integrantes por medio de la tripla (**V**, **E**, **C**), donde **V** denota el conjunto de individuos, **E** refiere a un grafo no dirigido cuyos ejes representan los vínculos entre los miembros del club, y **C** define la fuerza de estas relaciones —lo que se podría pensar como ponderaciones sobre los ejes de **E**—.

La investigación tuvo como enfoque central demostrar la capacidad del modelo para predecir la división del grupo por medio del algoritmo de *labeling* de ‘flujo máximo - corte mínimo’ de Ford y Fulkerson [3].

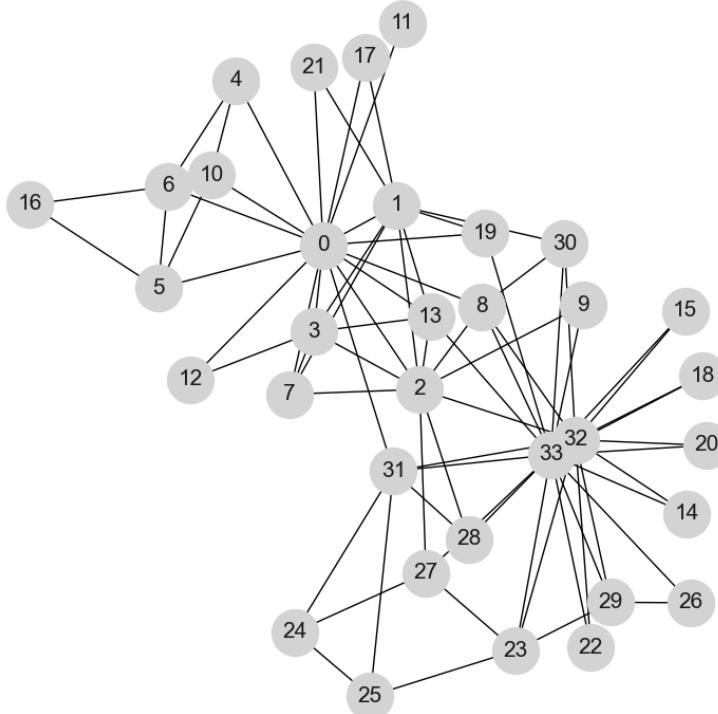


FIGURA 3. La red del *Club de Karate*. Cada nodo representa un individuo, cada eje la existencia de un vínculo por fuera del club.

En este análisis¹⁰ utilizaremos la representación matricial de **E** para evaluar la importancia de los distintos miembros en la red, y la matriz laplaciana asociada para evaluar el uso de autovectores como predictores de la división del grupo.

¹⁰El script asociado se puede encontrar en `./experimentos/club_de_karate.py`, los archivos con todos los resultados mencionados en `./experimentos/resultados/club-karate/`.

2.2. Centralidad de Autovector. La centralidad de autovector es una medida que se utiliza en el análisis de redes para evaluar la ‘importancia’ de los nodos que componen una red, relativa a la importancia de sus conexiones. Dada una matriz de conectividad $\mathbf{W} \in \mathbb{R}^{n \times n}$, se define:

$$(4) \quad \lambda x = \mathbf{W}x$$

donde λ es el autovalor en módulo máximo de \mathbf{W} y la coordenada x_i —del autovector x asociado a λ — denota la centralidad del nodo i , $\forall i : 1 \dots n$.

Intuitivamente, se puede pensar que la importancia de cada nodo es proporcional a la suma de las importancias de sus vecinos. Se puede demostrar [5] que, dado las características de esta matriz, el autovector asociado siempre será positivo en coordenadas.

En tanto la red de *Club de Karate*, podemos ver que la aplicación del método de la potencia sobre la matriz de conectividad asociada al grafo \mathbf{E} resulta en el siguiente autovector x^{11} :

0	0.355491	17	0.092400
1	0.265960	18	0.101403
2	0.317193	19	0.147913
3	0.211180	20	0.101403
4	0.075969	21	0.092400
5	0.079483	22	0.101403
6	0.079483	23	0.150119
7	0.170960	24	0.057052
8	0.227404	25	0.059206
9	0.102674	26	0.075579
10	0.075969	27	0.133477
11	0.052856	28	0.131078
12	0.084255	29	0.134961
13	0.226473	30	0.174758
14	0.101403	31	0.191034
15	0.101403	32	0.308644
16	0.023636	33	0.373363

FIGURA 4. Centralidad de autovector para la red del *Club de Karate*. La columna izquierda denota el nodo, la derecha su ‘importancia’.

Vemos que el nodo ‘0’ y el nodo ‘33’ son los más centrales. Esto no es casualidad, la red del *Club de Karate* está armada para tener a las dos figuras principales del conflicto en cada extremo —el instructor de karate y el presidente del club, respectivamente— para satisfacer la especificación del algoritmo de labeling que utiliza.

¹¹Notamos que el error relativo del resultado fue de $\approx 1.78e^{-14}$ en norma L_1 .

2.3. Autovectores de la matriz laplaciana. La matriz laplaciana $\mathbf{L} = \mathbf{D} - \mathbf{W}$ —donde \mathbf{D} es una matriz diagonal con elementos $d_{ii} = \sum_j w_{ij}$ y \mathbf{W} es una matriz de conectividad— sirve para medir distintas propiedades en una red.

En particular, el mínimo autovalor en módulo no nulo —llamado de conectividad algebráica, o valor de Fiedler— permite establecer un criterio sobre el que particionar la red en dos. El autovector asociado a este autovalor designará la pertenencia de un nodo a una u otra partición acorde a su signo.

Procedemos a analizar qué autovector de la matriz laplaciana asociada a \mathbf{E} permite predecir mejor la división que ocurrió en el *Club de Karate*. Para ello, calculamos los autovectores de la matriz con el método de la potencia con deflación¹² y medimos el valor absoluto de la correlación entre cada autovector y un vector que indica la división verdadera que ocurrió en el grupo.

18.136696	0.045692	3.013963	0.009556
17.055171	0.011443	2.749157	0.161363
13.306122	0.078413	2.487092	0.117379
10.921068	0.058283	2.0	0.0
9.777241	0.085078	2.0	0.0
6.996197	0.010429	2.0	0.0
6.515545	0.079173	2.0	0.0
6.331592	0.014177	2.0	0.0
5.618034	0.0	1.955050	0.011172
5.378595	0.244844	1.826055	0.068783
4.580793	0.079869	1.761899	0.027133
4.480008	0.044972	1.599283	0.05576
4.275877	0.010777	1.259404	0.000408
3.472187	0.000321	1.125011	0.333307
3.381966	0.0	0.909248	0.265918
3.376154	0.065159	0.468525	0.814727
3.242067	0.086678	0.0	nan ¹³

FIGURA 5. Correlación entre los autovectores asociados a los autovalores de la red del *Club de Karate* y un vector que indica la división verdadera que ocurrió en el grupo. La columna izquierda denota el autovalor, la derecha la correlación.

Vemos que el valor de conectividad algebráica —el autovalor mínimo en módulo no nulo— tiene asociado el mejor autovector para realizar el corte. En la figura (6.) se puede observar que sólo dos nodos obtuvieron una clasificación incorrecta: el ‘2’ y el ‘8’, lo que presenta un desempeño casi igual al logrado en la investigación original¹⁴.

¹²Notamos que el error relativo de los resultados tuvo una cota superior de $\approx 1e^{-13}$ en norma L_1 .

¹³El método de la potencia no está definido si el autovalor máximo en módulo es nulo. Este resultado da cuenta de ello. Como nos interesan solo los autovalores diferentes a ‘0’, optamos por descartarlo.

¹⁴En este sentido, solo el nodo ‘2’ recibió una clasificación incorrecta. Sobre el ‘8’, Zachary considera [6] que hubo un interés ‘egoísta’ que primó sobre la ‘afiliación’ política (es decir, las relaciones) del nodo.

0	0.112137	0	0
1	0.041288	0	0
2	-0.023219	1	0
3	0.055500	0	0
4	0.284605	0	0
5	0.323727	0	0
6	0.323727	0	0
7	0.052586	0	0
8	-0.051601	1	0
9	-0.092801	1	1
10	0.284605	0	0
11	0.210993	0	0
12	0.109461	0	0
13	0.014742	0	0
14	-0.162751	1	1
15	-0.162751	1	1
16	0.422765	0	0
17	0.100181	0	0
18	-0.162751	1	1
19	0.013637	0	0
20	-0.162751	1	1
21	0.100181	0	0
22	-0.162751	1	1
23	-0.155695	1	1
24	-0.153026	1	1
25	-0.160963	1	1
26	-0.187110	1	1
27	-0.127664	1	1
28	-0.095152	1	1
29	-0.167650	1	1
30	-0.073500	1	1
31	-0.098753	1	1
32	-0.130345	1	1
33	-0.118903	1	1

FIGURA 6. Autovector de Conectividad algebráica, la primer columna representa los nodos de la red, la segunda el autovector, la tercera la clasificación de cada nodo acorde a su signo y la cuarta la división verdadera que ocurrió.

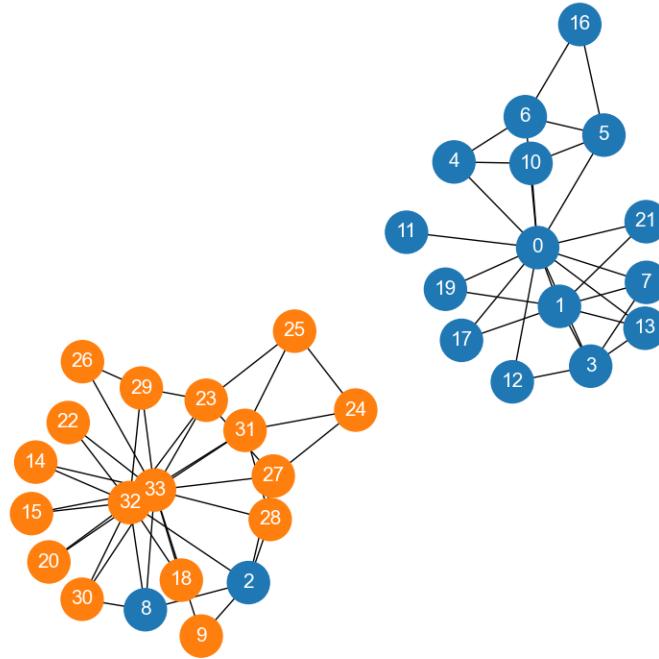
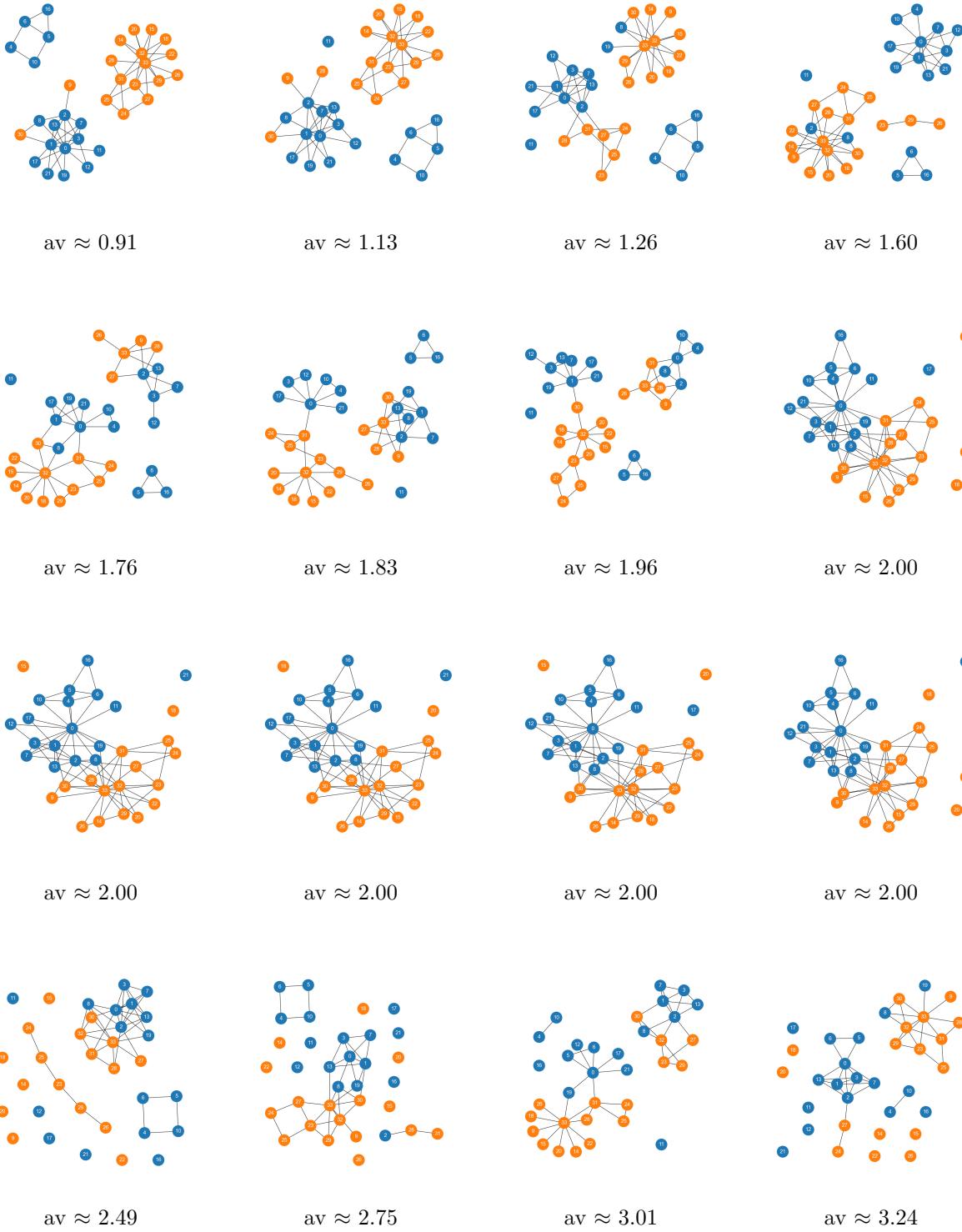


FIGURA 7. La división estimada de la red del *Club de Karate*. Los colores representan la división verdadera de la red. Azul: '0', Naranja: '1'.

Procedemos a detallar —a modo de comparación visual— los cortes realizados por los demás autovectores. Notamos que el proceso de corte sólo elimina los ejes existentes entre pares cuya clasificación difiere. Por ello, la cantidad de componentes conexos que puede generar un corte no es necesariamente dos.



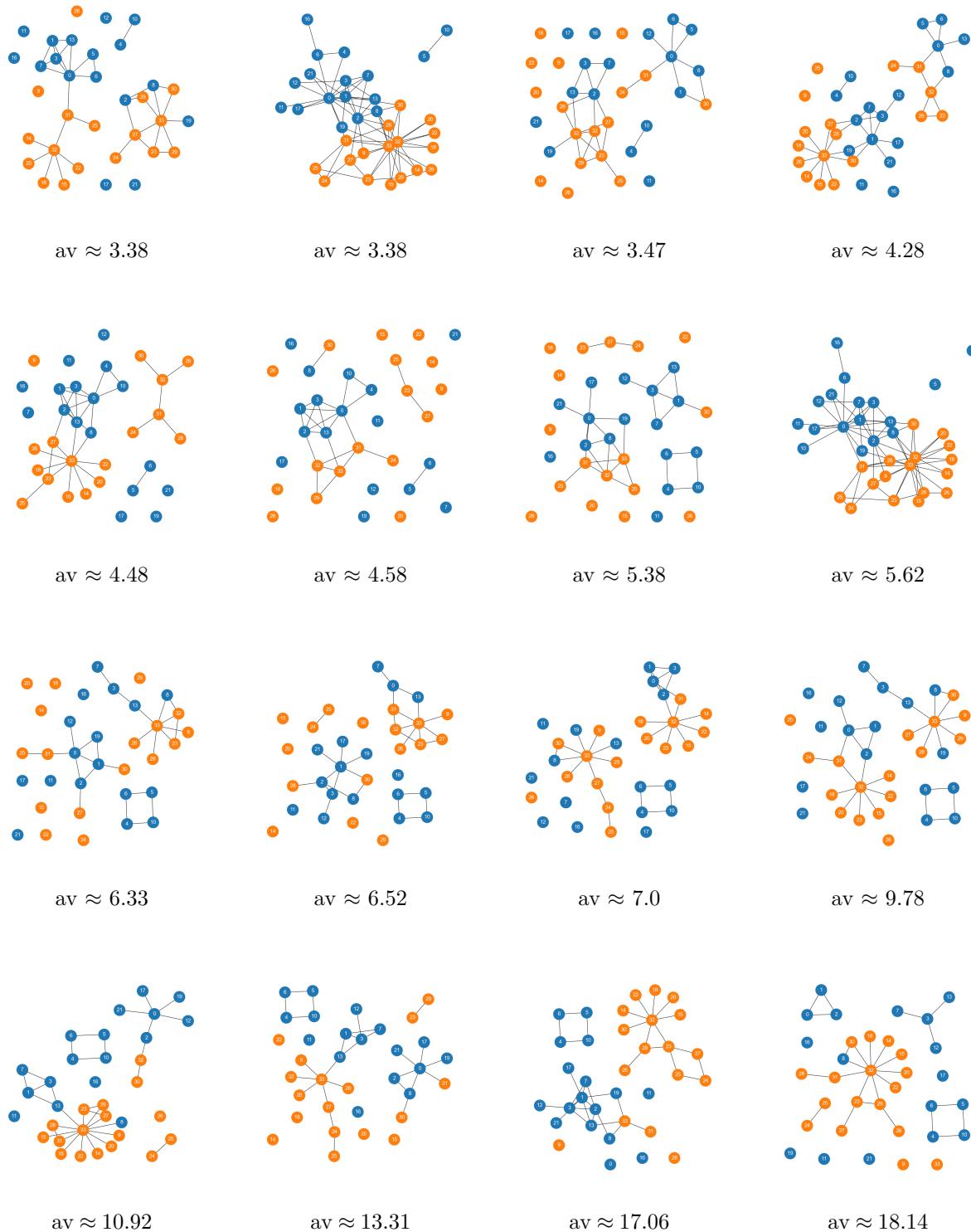


FIGURA 7. Todos los cortes posibles —salvo el de Fiedler y el nulo—, correspondientes a los autovectores asociados a la matriz laplaciana de la red del *Club de Karate*. Cada uno se designa por medio de su autovalor.

3. ANÁLISIS: RED ‘EGO’

3.1. Contexto. Una *ego-network* [4] es una red compuesta por las amistades que existen entre los amigos de un individuo, el ‘ego’. Estas redes son centrales para aplicaciones como Facebook, Google+ o Twitter.

En particular, dada una red ‘ego’, resulta de interés poder identificar los círculos sociales —conjuntos, disjuntos y anidados— a los que pertenece un usuario. Leskovec [4] propone un método de aprendizaje no supervisado para lograr inferirlos, que se nutre de la siguiente información: un grafo \mathbf{E} (la red) —donde se espera que exista una correlación fuerte entre un círculo y la densidad de conexiones entre los nodos que lo componen— y un conjunto de atributos \mathbf{C} , para cada nodo —donde se espera una correlación entre un círculo y la similaridad de los atributos de los nodos que lo componen—.

En este análisis estimaremos \mathbf{E} , una red ‘ego’ proveniente de Facebook, por medio de la construcción de una matriz de similaridad que utilice los atributos definidos en \mathbf{C} . También, buscaremos reducir su dimensionalidad por medio del análisis de componentes principales.

3.2. Matriz de similaridad. Queremos generar una aproximación de nuestra red ‘ego’ original \mathbf{E} en base a los atributos en \mathbf{C} de los usuarios que la componen. Es decir, buscamos un método que nos permita comparar los atributos de dos nodos diferentes y, bajo algún criterio propuesto, conectarlos o no en fin de replicar nuestra red verdadera.

Intuitivamente, una forma de adivinar si dos usuarios se encuentran conectados en una red es contando la cantidad de atributos que comparten. Se podría pensar que si coinciden en muchos existe una mayor probabilidad de que pertenezcan al mismo círculo, mientras que sino podría ser que ni siquiera se conozcan. Las estructuras que utilizaremos para replicar esta línea de pensamiento son las *matrices de similaridad*, las cuales dado un conjunto de datos X aplican una función a cada par de datos ij .

$$(5) \quad S_{ij} = f(X_i, X_j)$$

Nuestra tabla de atributos¹⁵ es tal que la primera columna contiene los tags correspondientes a cada usuario, y el resto de las columnas forman \mathbf{C} , donde la $fila_i(\mathbf{C})$ son los atributos del usuario tag_i . Tomemos entonces nuestra matriz de similaridad $\mathbf{S} = \mathbf{CC}^t$, tal que \mathbf{S}_{ij} expresa la similaridad entre la tag_i y tag_j computando el producto interno de sus respectivos atributos. Con esta información podemos luego proponer diferentes umbrales u entre el menor y mayor valor en \mathbf{S} , y establecer que $\mathbf{S}_{ij} > u$ indica que los tag_i y tag_j están conectados en nuestra aproximación.

Procedemos a mostrar los resultados obtenidos. Tomando $u \in [min\mathbf{S}_{ij}, max\mathbf{S}_{ij}] = [0, 22]$, $u \in \mathbb{Z}$, obtuvimos 22 diferentes aproximaciones de \mathbf{E} . Obsérvese que cuanto más alto sea el umbral, menores conexiones tendremos en nuestro grafo aproximado. En particular, con $u = 0$ se obtiene un grafo donde todos los usuarios están conectados entre sí, y con $u > 12$ grafos con ninguna arista. Nos enfocaremos entonces con los asociados a $u \in [0, 12]$, ya que son los que revelan información de interés.

¹⁵La misma puede encontrarse en [./catedra/ego-facebook.feat](#).

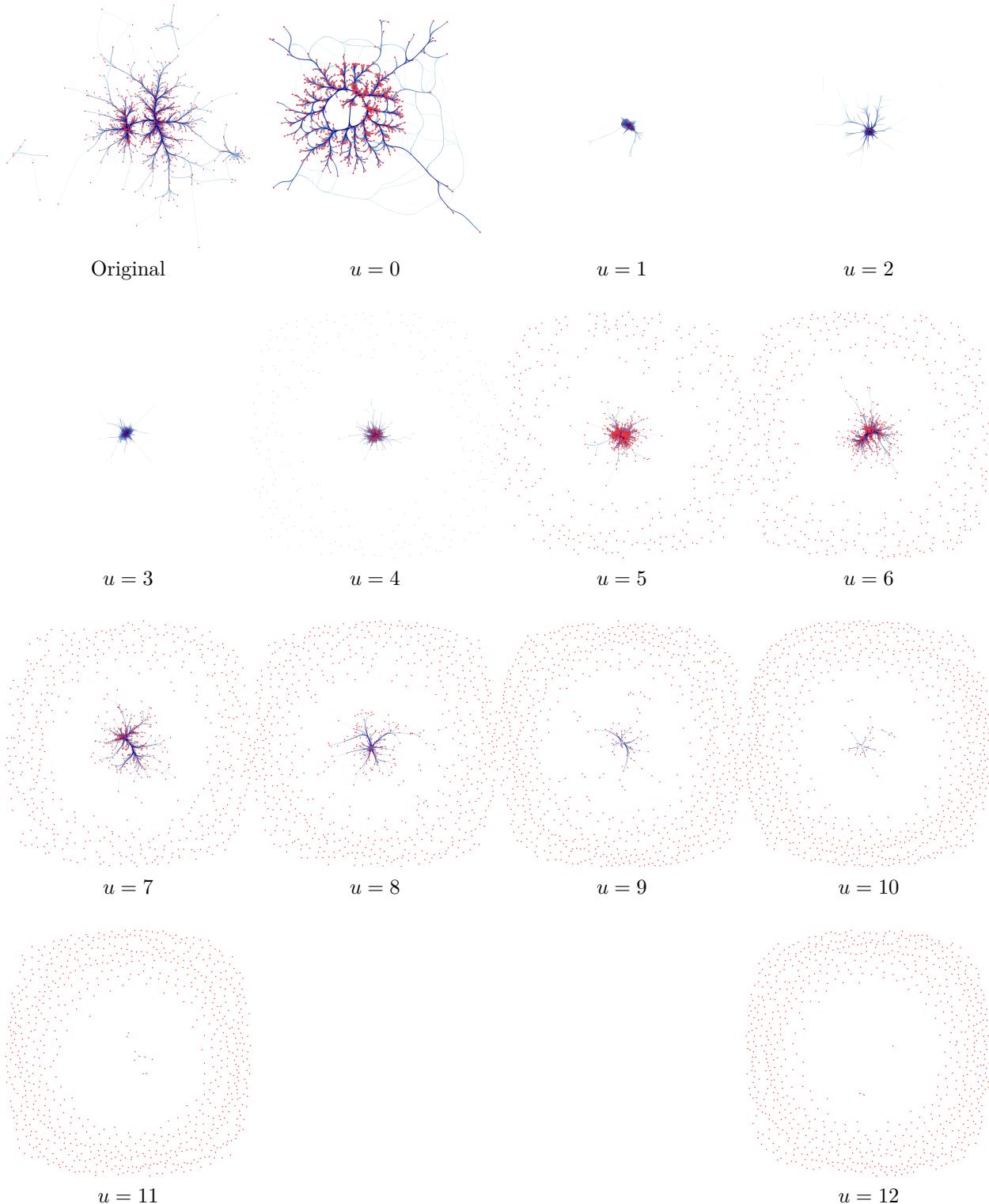


FIGURA 8. Todos los grafos correspondientes a los diferentes umbrales, tomados para la matriz de similaridad de los atributos **C**. Observar que al aumentar u crece la cantidad de nodos aislados.

3.3. Comparación con la red original. Nos encontramos con diferentes grafos construidos en base a los atributos en **C**. Buscamos ahora comparar nuestras aproximaciones con la red original. Un primer acercamiento a este problema fue calcular la cantidad de posiciones coincidentes entre las matrices de adyacencia y dividir por los elementos totales. La lógica por detrás del método es que nuestros grafos serán más similares entre sí por cada conexión y ‘no conexión’ acertada. Entonces, por cada 0 y 1 coincidente en nuestra aproximación y **E** nos acercaríamos más a un 100% de similitud.

0	% 7.6054231
1	% 32.947445
2	% 58.570142
3	% 70.993985
4	% 84.162733
5	% 91.537012
6	% 94.322073
7	% 95.214277
8	% 95.403337
9	% 95.446393
10	% 95.455457
11	% 95.458695
12	% 95.459342
13	% 95.459990

FIGURA 9. Similitud elemento a elemento de las matrices de adyacencia, la primer columna representa el umbral tomado para la aproximación y la segunda el porcentaje de elementos correctamente estimados con respecto a la original. Se incluye un grafo sin aristas, con $u = 13$.

Como puede observarse, desafortunadamente, los grafos con muy pocas conexiones, como prueban ser tomando los umbrales $u > 10$, tienen de los valores más altos, mientras que otros con cantidad de aristas similares a la red original, como con $u = 5$ o $u = 6$ (16.850 y 5.727 conexiones respectivamente), parecerían ser peores aproximaciones. Esto se da por la naturaleza rala de nuestras matrices de adyacencia. La de la red ‘ego’ **E** es de 786×786 , con 617.796 elementos, y tan solo 28.048 (% 4.54) de ellos son no nulos. Es decir, es rala en un % 95.46. Es por esto que comparar elemento a elemento no prueba ser un método muy descriptivo de qué tan buena es una aproximación, ya que cualquiera con pocos elementos coincidirá en un $\sim\%$ 95.

Empleamos entonces otros dos métodos de comparación: *la correlación de las matrices de adyacencia estiradas* y *la correlación de las listas de autovalores*. La correlación es una covarianza normalizada con valores en $(-1, 1)$, donde números mayores indican una mayor similitud. Esta es descrita por la siguiente fórmula:

$$(6) \quad \text{Corr}(x, y) = \frac{(x - \mu_x) \cdot (y - \mu_y)}{\sqrt{(x - \mu_x)^2 \cdot (y - \mu_y)^2}}$$

siendo μ el valor medio de los vectores.

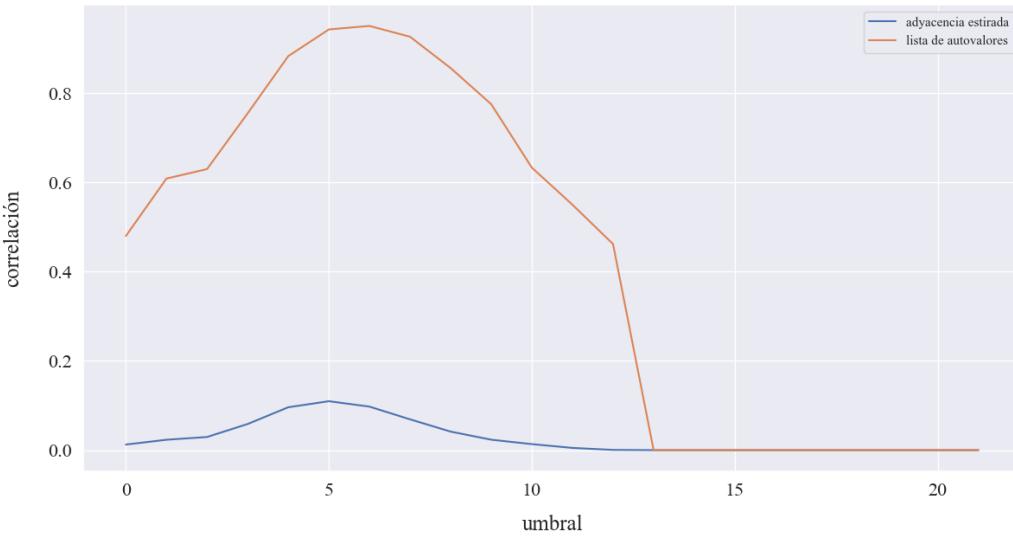


FIGURA 10. Correlaciones entre matrices de adyacencia y listas de autovalores según los diferentes umbrales.

En la figura (10.) se puede apreciar que se obtienen resultados considerablemente diferentes a los de nuestro primer método de comparación. En un primer lugar, siguiéndonos de la correlación entre las matrices de adyacencia, parecería que las mejores aproximaciones son las de cantidad de aristas similares a la red original, con los umbrales $u \in [4, 7]$, mientras que los grafos vacíos con umbrales más altos pasan a tener correlación 0. De la misma forma, analizando con los autovalores se obtienen resultados similares, donde la mayor correlación se halla con los umbrales de valores medios.

3.4. Optimización.

3.5. PCA.

4. CONCLUSIONES

5. APÉNDICE

REFERENCIAS

- [1] Gilles Brassard and Paul Bratley. *Fundamentals of Algorithmics*. 1995.
- [2] Richard L Burden and J Douglas Faires. *Numerical Analysis*. 2000.
- [3] L Ford and D Fulkerson. *Flows in networks*. Princeton University Press, 1962.
- [4] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.
- [5] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.
- [6] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.